UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

18|19

Trabajo fin de grado

Nuevos Esquemas de Aprendizaje para el Entrenamiento de Representaciones Agnósticas: Aplicación al Reconocimiento Facial

Marta Fernández de Barrio

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

Campus Internacional excelencia UAM CSIC+

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de la Telecomunicación

TRABAJO FIN DE GRADO

Nuevos Esquemas de Aprendizaje para el Entrenamiento de Representaciones Agnósticas: Aplicación al Reconocimiento Facial

Autor: Marta Fernández de Barrio
Tutor: Aythami Morales Moreno
Ponente: Julián Fiérrez Aguilar

mayo 2019

Marta Fernández de Barrio
Nuevos Esquemas de Aprendizaje para el Entrenamiento de Representaciones Agnósticas: Aplicación al Reconocimiento Facial

Marta Fernández de Barrio

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mi padre.

Es una locura odiar a todas las rosas sólo porque una te pinchó. Renunciar a todos tus sueños sólo porque uno de ellos no se cumplió.

El principito

Agradecimientos

En primer lugar deseo expresar mis más sinceros agradecimientos al grupo de investigación Biometrics and Data Pattern Analytics (BiDA-lab) y en concreto a mi tutor Aythami Morales por su enorme colaboración, esfuerzo, interés y ayuda indispensables para la realización de este proyecto. Gracias también por la confianza puesta en mi desde el primer momento. Agradecer de igual forma a Iván Bartolomé por su ayuda y compañía durante toda la elaboración del proyecto.

Agradecer también a mis padres y a mi hermana por el apoyo continuo, no solo en la realización del proyecto sino durante toda mi trayectoria académica. Hacer una mención especial a mi padre, el cual me ha ayudado durante toda mi vida académica mostrando una paciencia y ayuda de valor incalculable. Nunca hubiera llegado hasta aquí sin ellos.

Gracias a mis amigas de toda la vida por el apoyo y compresión ofrecidos durante estos cuatro años de universidad. Gracias también a mis compañeros de clase Berta Fernández, Andrea González y Manuel Moyano por acompañarme de forma incondicional durante estos años. Vuestro apoyo y ayuda han sido de vital importancia para llegar hasta aquí.

Por último, mis más sinceros agradecimientos a Santiago Palmero, no solo por la ayuda académica ofrecida, si no también por su increíble paciencia, apoyo y ánimo sin los cuales no puedo imaginarme estos últimos años.

Resumen

Este trabajo propone el objetivo de eliminar información sensible en los procesos de toma de decisiones de los algoritmos de aprendizaje profundo. Para ello se plantea una nueva representación agnóstica de los vectores de características de las redes neuronales, de forma que estos dificulten a las redes la clasificación de los grupos protegidos. Una creciente aparición de estudios demuestra la existencia de un gran riesgo de efectos discriminatorios producidos por los algoritmos actuales. Esta situación ha conllevado la aparición de leyes que obligan a eliminar la posibilidad de cualquier tipo de discriminación algorítmica. Por todo ello este trabajo está motivado a la creación de un nuevo modelo capaz de cumplir con las necesidades algorítmicas requeridas hoy en día. Se parte de un modelo basado en la generalización de triple loss que, no solo optimice el rendimiento de verificación, sino que también disponga de un proceso de eliminación de sesgo. De forma más concreta el proyecto se centra en la eliminación de etnia y género. El modelo es evaluado mediante los algoritmos más punteros de la actualidad y el uso bases de datos disponibles al público. De forma adicional se ha generado una nueva base da datos balanceada entre 2 clases de género y 3 clases de etnia, la cual ha facilitado notablemente las tareas de eliminación de sesgo. El conjunto dispone de más de 20K identidades que resultan en más de 100K imágenes. Todas ellas muestran variedades de pose, calidad, iluminación, entre otras. Los resultados de las distintas pruebas han demostrado la posibilidad de reducción de una gran cantidad de información sensible al mismo tiempo que se mantiene un alto rendimiento en verificación.

Palabras clave

Reconocimiento facial, información sensible, representación agnóstica, discriminación algorítmica, toma de decisiones justa.

Abstract

This work aims at eliminating sensitive information in the decision-making processes of deep learning algorithms. For this purpose, a new agnostic representation of embedding vectors of neural networks is proposed, in such a way that they make it difficult for networks to classify the protected groups. A growing number of studies prove the existence of a great risk of discriminatory effects produced by current algorithms. This has led to the emergence of laws which require the elimination of the possibility of any type of algorithmic discrimination. Therefore, this work is intended to create a new model capable of meeting the algorithmic needs required today. It starts from a model based on the generalization of triple loss, which not only optimizes the verification performance, but also has a process of bias elimination. More specifically, the project focuses on the removal of ethnicity and gender. The model is evaluated using cutting-edge algorithms and publicly available databases. In addition, a new balanced database has been generated, which has two classes of gender and three of ethnicity, notably facilitating bias removal tasks. The set has more than 20K identities resulting in more than 100K images. All of them show varieties of pose, quality, and lighting, among others. The results of the experimentation have shown the possibility of reducing a large amount of sensitive information, while maintaining a high verification performance.

Keywords

Face recognition, sensitive information, agnostic representation, algorithmic discrimination, fair decision-making.

Índice

Listas

Lista de ecuaciones

Lista de figuras

Lista de tablas

1

Introducción

Actualmente los sistemas de reconocimiento facial han asumido un papel de liderazgo debido al buen rendimiento en las tareas de reconocimiento humano. Hasta ahora las investigaciones para desarrollar algoritmos de reconocimiento han centrado su atención en adquirir un porcentaje de acierto lo más alto posible. En este sentido, se han hecho grandes avances que han permitido su despliegue en aplicaciones reales. No obstante, actualmente se trabaja en nuevas líneas de investigación más allá del rendimiento, como es la privacidad y el derecho a la no discriminación de cualquier ciudadano. En este contexto se entiende como discriminación a la variación de las decisiones del algoritmo en función de la pertenencia de una persona a un grupo específico como, por ejemplo, la etnia, el género o incluso la edad. Dichos grupos se conocen como grupos protegidos o vulnerables [1].

El uso de perfiles algorítmicos, que contengan información sobre la pertenencia a un grupo vulnerable es, en cierto sentido, inherentemente discriminatorio. Si dichos sistemas se limitaran únicamente a los laboratorios de investigación de la inteligencia artificial, la preocupación sobre su comportamiento sería exponencialmente menor. Sin embargo, los algoritmos de reconocimiento biométrico están cada vez más expandidos, es decir, cada vez es más normal su existencia en el día a día de la sociedad. Tanto es esto que determinadas decisiones importantes sobre las personas han puesto su confianza en este tipo de tecnologías. Algunas de ellas incluyen la emisión de prestamos, admisión de estudiantes, temas jurídicos o incluso la provisión de atención médica [2]. Rápidamente se puede observar que una decisión injusta en alguno de estos ámbitos puede condicionar gravemente la vida de una persona. Por ejemplo, una persona puede ser acusada de un crimen debido a una identificación errónea de una cámara de seguridad [3]. Por ello, si bien el deep learning consta de un potencial increíble, es importante no sólo valorar sus fortalezas, sino también ser consciente de sus debilidades [4].

En esta línea, se han realizado numerosas investigaciones para observar los efectos negativos que pueden causar este tipo de tecnologías. Una investigación que se realizó en numerosos departamentos de policía demostró que las personas de color tienen una probabilidad mayor de ser detenidas y sometidas a registros de reconocimiento facial. Esta situación, entre otras, supone una amenaza para las libertades civiles. De hecho, diversos estudios han demostrado que los algoritmos de reconocimiento facial reconocen erróneamente en un alto porcentaje de error a las personas de color, de género femenino o a los jóvenes [3].

Por todo ello, las nuevas tecnologías biométricas y la privacidad han sido dos conceptos muy enfrentados a lo largo de las últimas décadas. Los gobiernos son conscientes de este problema y de la importante necesidad de un compromiso entre los niveles de ambos conceptos. En consecuencia, la protección de datos ha conseguido un gran protagonismo en nuestra sociedad [5]. Asimismo, la biometría ha abierto un nuevo debate en el que se han involucrado personalidades de diversos ámbitos, como la filosofía, la abogacía o la legislatura, entre otros. Uno de los principales motivos de esta discusión es la definición de justicia en el término tecnológico, el cual debe ser expresado en términos matemáticos [6]. Este hecho ha llevado a diversas definiciones que, en muchas ocasiones, han resultado ser contradictorias entre ellas.

Por este motivo, entre otros, en abril de 2016 se aprobó el Reglamento General de Protección de Datos (RGPD) el cuál regula la recolección, el almacenamiento y el uso de la información personal. El articulo 22: *"Decisiones individuales automatizadas, incluida la elaboración de perfiles"* preocupa especialmente a los grupos de machine learning, ya que resulta en la prohibición de una amplia gama de algoritmos. Este articulo dice lo siguiente: *"Todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar"*. Asimismo el articulo 9: *"Tratamiento de categorías especiales de datos personales"* prohíbe el uso de datos que clasifiquen a las personas en grupos vulnerables a la discriminación [7]. Esto produjo que se incrementara de forma exponencial la necesidad apremiante de algoritmos eficaces que pudieran funcionar dentro de este nuevo marco jurídico.

Además de los derechos establecidos en el RGPD, las personas tienen el derecho a la no discriminación y por tanto a la no discriminación algorítmica. Por ende, el artículo 21 de la *Carta de los derechos fundamentales de la Unión Europea* establece lo siguiente: *"Se prohíbe toda discriminación, y en particular la ejercida por razón de sexo, raza, color, orígenes étnicos o sociales, características genéticas, lengua, religión o convicciones, opiniones políticas o de cualquier otro tipo, pertenencia a una minoría nacional, patrimonio, nacimiento, discapacidad, edad u orientación sexual"*. Asimismo el artículo 14 del *Convenio Europeo de Derechos Humanos* dice lo siguiente: *"El goce de los derechos y libertades reconocidos en el presente Convenio ha de ser asegurado sin distinción alguna, especialmente por razones de sexo, raza, color, lengua, religión, opiniones políticas u otras, origen nacional o social, pertenencia a una minoría nacional, fortuna, nacimiento o cualquier otra situación."* Por último, también es necesario destacar los artículos del 18 al 25 del *Tratado de funcionamiento de la Unión Europea* los cuales forman un conjunto de artículos que recogen los derechos sobre la no discriminación y la ciudadanía de la unión [7].

Se puede observar fácilmente que las leyes avalan los derechos a la no discriminación. Sin embargo, eliminar la discriminación en términos tecnológicos no es tan sencillo. Cuando un modelo de deep learning ha sido entrenado, no siempre se conoce como se ha tomado esa decisión. En esta línea, por lo tanto, se puede decir que existe una carencia de sentido común. A pesar de que los modelos

de deep learning tienen una alta eficacia en la percepción de patrones, estos no están diseñados para entender el significado de dichos patrones y por lo tanto no existe un razonamiento sobre ellos. Esto da lugar a situaciones comprometidas donde no se podría dar a una persona la explicación de la decisión tomada. Por ejemplo, si un banco utiliza estos algoritmos para evaluar la concesión de un crédito y la respuesta es negativa, el banco no será capaz de dar una explicación clara de por qué se le negó préstamo [4]. En esta situación, no solo se estaría negando algo ético, sino que también se estaría incumpliendo el *derecho a la explicación* recogido de nuevo en el RGPD. Entre ellos destacan el artículo 13: *"Información que deberá facilitarse cuando los datos personales se obtengan del interesado"* y el artículo 14: *"Información que deberá facilitarse cuando los datos personales no se hayan obtenido del interesado"* que señalan lo siguiente: *"La existencia de decisiones automatizas, incluida la elaboración de perfiles, a que se refiere el artículo 22, apartados 1 y 4, y, al menos en tales casos, información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado."* [7]. Por consiguiente, se puede observar la necesidad inminente del desarrollo de algoritmos que faciliten una máxima transferencia en la toma de decisiones.

Existen dos formas principales en las que se pueden producir decisiones influenciadas por el sesgo [6]:

- Los datos recopilados no representan de forma fiel la sociedad actual.
- El algoritmo ha heredado prejuicios existentes en la sociedad.

El primer caso ocurre cuando un algoritmo es entrenado con una base de datos no realista, es decir, que no representa por igual a todos los grupos sociales que existen actualmente. Esta discriminación existente en los datos, se manifiesta en los resultados del reconocimiento automático, ya que este tendrá una tasa de error considerablemente mayor cuando se trate de reconocer a los miembros de los grupos que no han sido bien representados, existiendo una mayor incertidumbre asociada con esas predicciones. Por ejemplo, si los datos están formados en su mayoría por personas de piel clara, el sistema resultará inevitablemente peor para reconocer rostros de piel oscura [6].

Los grandes conjuntos de datos disponibles de forma pública frecuentemente están compuestos por rostros de celebridades. Como es de esperar, estos conjuntos están lejos de representar un mundo real. De hecho, las mujeres famosas tienden a ser más jóvenes que el género masculino, lo cual introduce un claro desbalanceo de edad [8]. Diversas bases de datos que son muy utilizadas en este ámbito tienen este problema. Entre ellas se encuentran bases de datos como VGGFace2, Ms-celeb-1m, Labeled faces in the wild (LFW) o MegaFace, entre otras [8].

El segundo caso ocurre cuando los sistemas son entrenados para tomar decisiones basadas en datos históricos, ya que estos heredarán, naturalmente, los sesgos del pasado. Los autores Solon Barocas y Andrew D. Selbst explican que el aprendizaje automático depende de los datos que se han recogido de la sociedad, y en la medida en que la sociedad contenga desigualdad, exclusión u otros rastros de discriminación, también lo harán los datos. Asimismo, señalan textualmente: "La

dependencia irreflexiva de la minería de datos puede negar a los miembros de los grupos vulnerables la plena participación en la sociedad" [9].

Por ejemplo, en el pasado, las mujeres solían aparecer representadas en las cocinas de forma frecuente, lo que pasaba de forma contraria para el género masculino. Por ello, un sistema de clasificación de generó aprendió que la cocina era una condición clave para identificar a una persona del género femenino. Si bien es cierto que un humano, en general, nunca tendría en cuenta en que lugar de la casa se encontrara el sujeto, para el algoritmo condicionaba su decisión, lo cual es absolutamente discriminatorio [8].

Con bases de datos que contienen conjuntos de datos relativamente pequeños es más fácil identificar las correlaciones entre las variables que introducen el sesgo, también conocidas como variables sensibles, y las variables no sensibles. No obstante, como cabe esperar, la reducción o eliminación de la información correlacionada con los datos sensibles puede resultar en un mal funcionamiento del sistema y por tanto imposibilitar su uso [7]. En esta línea, Toon Calders y Sicco Verwer mencionan el siguiente ejemplo: *"El código postal puede revelar información racial y, al mismo tiempo, ofrecer información útil y no discriminatoria sobre el incumplimiento de los préstamos"* [10].

De forma contraria, a medida que el volumen de los conjuntos de datos aumenta, las correlaciones pueden volverse cada vez más complejas y difíciles de detectar. Por ejemplo, la relación entre la geografía y los ingresos puede ser clave, pero es probable que existan correlaciones como, por ejemplo, entre la geografía y la etnia, dando lugar a efectos discriminatorios [7].

Por ende se puede concluir que los errores comentados, a pesar de parecer irrelevantes, tienen importantes consecuencias negativas en la toma de decisiones de los sistemas automáticos. La introducción de sesgo no es fácilmente identificable e, incluso cuando se conoce el problema, este se convierte en un auténtico reto para eliminarlo.

1.1. Introducción al reconocimiento facial

El objetivo del reconocimiento facial es extraer la mayor cantidad de información posible, como la identificación, la pose, el sexo, la edad etc., de un rostro. De acuerdo a [11], se obtienen los siguientes puntos sobre estos algoritmos.

Uno de los requisitos más desafiantes hasta el momento ha sido que los detectores faciales detecten rostros con distintas características como la pose o la iluminación, entre otras problemáticas. Dicho reto ha sido mitigado mediante el uso del deep learning o aprendizaje profundo y más concretamente con el uso de las redes neurales convolucionales profundas (DCNN). En este sentido, las bases de datos contienen diversas características, lo cual permite a las redes ser más robustas a dichas variaciones.

Una simple imagen facial contiene información no solo sobre quien es esa persona, si no que proporciona también información sobre el género, la etnia o la edad. Dichos datos se conocen en términos tecnológicos como biometría suave. En general, identificar dichos datos es una tarea fácil para los humanos. Cuando una persona mira una cara en una imagen sabe rápidamente la edad, la pose, el sexo, expresiones, etc. No obstante, cuando las máquinas están diseñadas para realizar estas tareas, a menudo se construyen algoritmos independientes para resolver cada una de ellas. Este proceso se conoce como aprendizaje multitarea (MTL).

Durante las últimas décadas el MTL estaba muy limitado. Esto se debía a que cada tarea necesitaba representaciones de características diferentes. Por ejemplo, la detección de caras utilizaba el descriptor HOG, mientras que el reconocimiento utilizaba del descriptor LBP. Asimismo, este problema también existía en la clasificación de atributos como la edad o el género. Sin embargo, gracias a la aparición del deep learning, surgió la posibilidad de diseñar una red profunda que llevara a cabo todas estas tareas simultáneamente, compartiendo las características y explotando las relaciones entre ellas. De esta forma, se obtiene una solución robusta a la vez que se reduce el overfitting o sobreentrenamiento. Por tanto, los recientes métodos basados en deep learning para la clasificación de los atributos han resultado en impresionantes resultados, identificándolos con una precisión de más del 95 % [5].

Algunos de los atributos faciales comentados pueden utilizarse para mejorar el rendimiento de estos sistemas. Recientemente, un método basado en la identificación de atributos muestra que un gran número de estos por sí solos pueden dar buenos resultados. Dichos atributos son rasgos semánticos que resultan más fáciles de aprender que las identidades faciales. Este método resulta muy eficaz en la autenticación y más concretamente en teléfonos móviles.

El aprendizaje profundo se ha convertido en una técnica esencial para el reconocimiento facial. Estos sistemas generalmente se basan en tres procesos para cumplir su función. En todos ellos se puede observar la importancia de las redes de aprendizaje profundo. Dichos procesos se indican a continuación.

Detección de rostros

De forma contraria a la detección genérica de objetos, la detección facial es más desafiante debido al amplio rango de variaciones que pueden tener los rostros. Este proceso desempeña un papel crucial en un proceso de reconocimiento facial y constituye el primer paso en este tipo de sistemas. Como se ha comentado, el uso de las DCNN es básico en este proceso. Los métodos de detección facial se pueden dividir en dos subcategorías:

Region-Based:

Se basa en la generación de distintas imágenes como propuestas y mediante el uso de las DCNN se detecta si estas contienen un rostro o no. Un inconveniente de este método es que las caras difíciles de detectar seguirán siendo difíciles en cualquier propuesta. Además, como cabe esperar, la

generación de todas las posibilidades requiere un tiempo de computación demasiado alto.

Sliding-window based:

Se trata de un método basado en ventanas deslizantes donde se calcula en cada una de ellas una puntuación obtenida a partir de un mapa de características. Dicha puntuación se va obteniendo a diferentes escalas en cada ubicación de la imagen. Puede ser implementado utilizando solo una operación de convolución, por lo que se considera que este método es computacionalmente más rápido que el anterior.

Detección de puntos característicos

La detección de puntos clave faciales es también un componente importante de pre-procesamiento para el reconocimiento facial y las tareas de verificación.

En este proceso, el volumen de la base de datos utilizada puede ser crucial. Un conjunto de datos a gran escala hará que el sistema sea más robusto frente a desafíos como la pose, la iluminación, el tamaño o la calidad de la imagen. Las capas más profundas de las redes serán las encargadas de codificar este tipo de información más abstracta.

La mayoría de los algoritmos de localización de puntos característicos faciales utilizan dos enfoques principales.

Model based:

Este proceso se basa en modelar un algoritmo durante el entrenamiento y adaptarlo a nuevas caras durante la fase de test. Hasta ahora estos modelos se han basado en las formas bidimensionales. No obstante, actualmente se han desarrollado modelos basados en las formas tridimensionales (3-D).

Cascaded regression based:

La transformación del rostro en un vector de características se puede considerar, naturalmente, un problema de regresión. Por esta razón, a lo largo de los últimos años se han propuesto multitud de modelos basados en esta técnica. Por lo general estos métodos mapean directamente la apariencia de la imagen al resultado final. A pesar de ello, el rendimiento de este tipo de métodos depende de la robustez de los descriptores. Numerosos estudios han hecho uso del deep learning para diseñar estos métodos. Por ejemplo, en [12], propusieron una cascada de redes DCNN cuidadosamente diseñadas en las que, en cada nivel, las salidas de las redes se fusionan para estimar los puntos de referencia y lograr un buen rendimiento. En [13] se hace uso de una única red DCNN, para proporcionar un descriptor de puntos característicos único.

Identificación y verificación

Existen dos componentes principales en los procesos de identificación y verificación de rostros. Por un lado se necesita una representación facial robusta y por otro lado un modelo de clasificación

(identificación) o una medida de similitud (verificación).

Para las tareas de identificación y verificación faciales, es importante obtener características de identidad discriminatorias y robustas. En el proceso de verificación cada una de las imágenes de la cara es modelada a través de una DCNN para obtener un vector de características. Una vez obtenido, se calcula el grado de similitud entre ambas imágenes. Para ello las medidas más comunes son la distancia de norma L2 y la similitud coseno. Para la tarea de identificación, de forma similar a la tarea anterior, se pasan las imágenes de la cara a través de las DCNNs. Se obtienen los vectores de características correspondientes y se almacenan en la base de datos. Cuando se proporciona una nueva imagen para ser verificada, se obtiene su vector correspondiente y se calcula su similitud con toda la información almacenada.

Sin embargo, la eficacia del método también depende en gran medida de la calidad de los rostros detectados. Además, se han desarrollado varios métodos MTL para la detección, que implican el entrenamiento simultáneo de la tarea de detección de rostros junto con una tarea correlacionada como la estimación de puntos clave faciales. Dicha estimación puede ayudar a la red a determinar la estructura de la cara.

1.2. Objetivos

Como objetivos principales de este trabajo se destacan dos. Por un lado, la generación de una base de datos (DiveFace) con una distribución uniforme de género y etnia (Personas de color, asiáticos y caucásicos). Este conjunto de datos será el utilizado durante todos los entrenamientos, ya que evitará los problemas que acarrean las bases de datos desbalanceadas. Por otro lado, se busca el análisis de nuevas representaciones de los vectores de características usados en algoritmos de reconocimiento facial, que permitan eliminar la información sensible en los procesos de toma de decisiones. Dichas representaciones serán compatibles con los modelos pre-entrenados ya existentes. Entre ellos encontramos ResNet-50 y VGGFace. Por ello, los experimentos realizados en este proyecto contendrán experimentos sobre ambas arquitecturas, en los cuales se evaluarán las dificultades causadas por cada una de ellas.

Se realizará una evaluación de distintas estrategias de aprendizaje de los algoritmos que permitan conseguir los objetivos mencionados anteriormente. Más concretamente, en primer lugar se estudiarán las estrategias de forma independiente tanto para etnia como para género. Posteriormente se realizará un análisis del funcionamiento de los distintos métodos combinados.

1.3. Motivación

Por todo lo comentado hasta el momento se puede observar que existe una clara necesidad de algoritmos que no favorezcan o desfavorezcan a determinados grupos vulnerables de la sociedad. Como ya se ha visto, los procesos de toma de decisión de este tipo de métodos cada vez tienen más influencia en la vida humana. Ámbitos económicos, educativos, de la salud, son solo algunos de los ambientes en donde los sistemas de aprendizaje automático ya están plenamente instaurados. La constante aparición de leyes que protegen a las personas de la discriminación algorítmica indica una preocupación social clara. No obstante, a pesar de la existencia de los reglamentos comentados, existe un urgente reclamo de modelos que sean incapaces de discriminar a las personas individuales o a determinados grupos. Por ello, este trabajo basa su motivación en la creación de dichos algoritmos, los cuales sean capaces de asegurar a la sociedad que las decisiones tomadas de forma automática no basarán su decisión en la pertenencia a un determinado colectivo. Así, se fomentará la creación de una población sin racismo o machismo en el mundo tecnológico, el cual es cada vez más determinante en la vida de las personas.

1.4. Organización

Este proyecto se organiza de la siguiente forma. En primer lugar se realiza una exposición del estado del arte en el ámbito de la discriminación algorítmica en reconocimiento facial. A continuación se encuentran los apartados de diseño y desarrollo donde se explica todo el proceso seguido. Tras ellos se presenta un apartado de resultados donde se expone todo lo obtenido tras el proceso. Por último se finaliza con una conclusión de todo lo comentado en el proyecto y una exposición del trabajo futuro.

2

Estado del arte

El estudio de diferentes técnicas para eliminar la discriminación de los sistemas biométricos es un ámbito que ha sido investigado a lo largo de los últimos años y, por tanto, existen diversas investigaciones de diferentes comunidades de machine learning. En entre ellas destacan las investigaciones realizadas en [1], [14], [15], [2], [16], [10], [17], [3], [18], [19] y [8].

En [14] y en [15] se estudia el aprendizaje de nuevas representaciones con el objetivo de mejorar la justicia algorítmica. En [14] se presenta un nuevo método para entrenar redes neuronales justas con el objetivo de proporcionar una mejor justicia individual y de grupo. Se plantea un nuevo enfoque denominado Gradient Reversal Against Discrimination (GRAD) que utiliza una red que predice una clase y un atributo sensible. Sin embargo, las predicciones de dicho atributo se invierten antes de la actualización de pesos, haciendo que la red sea incapaz de predecir la información protegida. En [15] se estudia un clasificador capaz de tomar una decisión de una manera que sea totalmente agnóstica a una información protegida dada, incluso si esta se trata de información contextual.

De forma similar a [14], [2] abarca la justicia individual y de grupo como un problema de optimización con el objetivo de conseguir el mejor rendimiento posible mientras se oculta cualquier información sobre la pertenencia a un grupo protegido. En [16] se analiza cómo limpiar los conjuntos de datos de tal manera que puedan extraerse reglas de clasificación justas, pero no decisiones discriminatorias basadas en datos sensibles. Para ello se propone un método basado con la transformación de los datos que puedan estar correlacionados con la información sensible. En la misma linea, en [10] se realiza una investigación sobre la modificación del modelo de Bayes con el objetivo de que este clasifique de forma independiente a los atributos sensibles.

Un marco de aprendizaje automático para incorporar la equidad en los clasificadores probabilísticos se plantea en [17], donde se propone modificar la distribución de probabilidades de los clasificadores. En [18] se propone un diseño de investigación, centrado en las plataformas web, análogo a los estudios de auditoría, los cuales son típicamente experimentos en los que los investigadores examinan un proceso social con el fin de detectar una posible discriminación.

Una evaluación del rendimiento de la clasificación de género basada en el rostro se realiza en [3]. En este se evalúa el sesgo presente en los algoritmos de análisis facial y los conjuntos de datos con

respecto a los subgrupos fenotípicos. Los resultados efectivamente muestran que las mujeres de piel oscura son clasificadas con tasas de error notablemente mayores que las de los hombres blancos.

Un estudio destacado es [19]. En este se propone el uso de una red neuronal convolucional multitarea (MTCNN) que emplea un ajuste dinámico de la función de perdidas dirigido a clasificar de forma conjunta el género, la edad y la etnia, así como a mitigar los sesgos relacionados. Para ello se propone modelar la arquitectura FaceNet. Se plantea que la red MTCNN aprenda las relaciones entre las tareas de clasificación a partir de los datos y por tanto que sea capaz de crear un esquema de ponderación dinámica que asigne automáticamente los pesos de la función de perdidas para cada tarea durante el entrenamiento. De este modo, la MTCNN debe asignar un peso mayor para una tarea no relevante con una pérdida menor, con el fin de reducir la pérdida total.

Este trabajo experimenta el efecto del método propuesto en la clasificación de género, de etnia y de edad. Los resultados muestran que el modelo mejora el rendimiento que conseguía FaceNet previamente. Esto se debe a que la existencia del sesgo decrementaba el rendimiento de la clasificación en determinados grupos y por tanto, al ser el sesgo reducido, la precisión del sistema mejora notablemente.

Otra investigación a destacar es [8]. Esta introduce un algoritmo que asegure que la red no reconozca un sesgo conocido en la base de datos y que elimine las variaciones espurias de la representación de características de una tarea de clasificación primaria. De esta forma se propone un algoritmo de aprendizaje y desaprendizaje con el objetivo de que aprenda una tarea principal mientras que, al mismo tiempo, desaprende las variaciones espurias.

Este trabajo realiza 3 experimentos principales: 1) Eliminación del sesgo de una red, 2) eliminación de un sesgo extremo de una red y 3) eliminación simultánea de múltiples variaciones espurias. En los 3 casos, se crea una primera red que solo se entrene para la tarea principal y una segunda red que se entrene tanto para la tarea primaria como para la eliminación del sesgo.

Para los dos primeros casos se utiliza la base de datos IMBD la cual contiene sesgo en edad entre hombres y mujeres (las mujeres tienden a ser más jóvenes). No obstante, en el segundo caso, la base de datos se modifica para que el sesgo sea mayor. En ambos experimentos se demuestra que haciendo uso de la primera red, la clasificación de la edad está influenciada por el género. De forma contraria, el uso de la segunda red elimina esta relación haciendo que el rendimiento del sistema aumente. En el último caso se analiza el efecto de eliminar la información de género, edad, pose y origen, dejando en cada experimento una de ellas como tarea primaria. En el resultado se obtiene que eliminando más del 90 % de información de la tarea secundaria, el rendimiento de la primaria prácticamente no se ve afectado.

3

Diseño

Este proyecto parte del método de aprendizaje propuesto en [1] y cuyo código fue suministrado por el grupo *Biometrics and Data Pattern Analytics Lab* al inicio del trabajo. El objetivo de este proyecto se trata de entrenar una nueva representación, que, manteniendo el potencial discriminante para identidad, no contenga información asociada al género o la etnia. La Figura 3.1 ilustra el objetivo básico de este estudio.

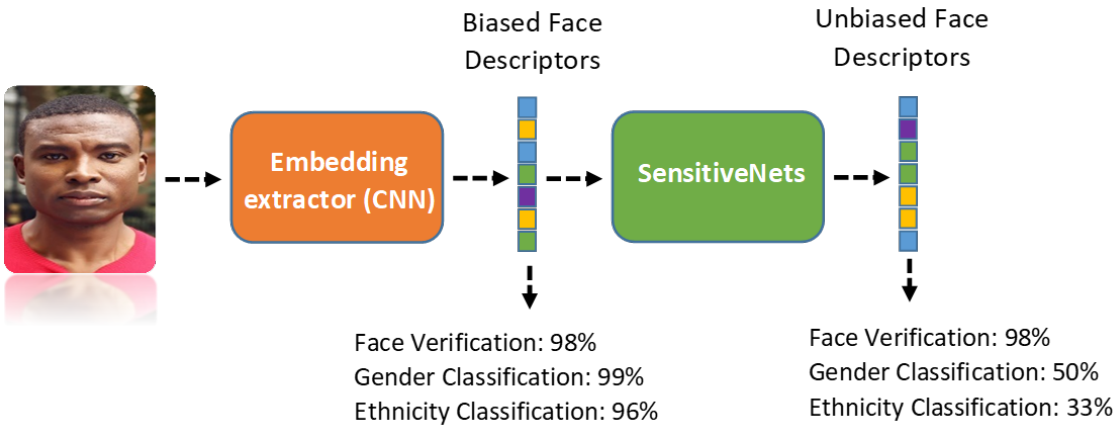


Figura 3.1: Ilustración de la finalidad del proyecto. Se puede observar como, sin una perdida de rendimiento, la red es incapaz de detectar etnia o género. [1]

Para cumplir este objetivo, el método opta por usar un algoritmo basado en triplet loss. La versión original de esta técnica tiene como finalidad optimizar el reconocimiento. No obstante, en este proyecto se propone una nueva versión que cumple también la meta de cegar a la red la información sensible. A continuación se presenta el modelo propuesto.

En primer lugar, se parte de una arquitectura pre-entrenada que da lugar a un vector de características por cada imagen. De forma matemática este vector se define como $X \in R^d$ y contiene el sesgo a eliminar. El funcionamiento de triplet loss se basa en la generación de tríos de imágenes. Estos conjuntos están compuestos por tres imágenes diferentes de dos personas distintas. Las imágenes son las indicadas a continuación. Cada conjunto de imágenes generado será parte de una lista de triplets denominada T. En la Figura 3.2 podemos ver un ejemplo de su estructura.

- Imagen Anchor (A): Imagen que pertenece al mismo usuario de la imagen positiva.
- Imagen Positiva (P): Imagen que pertenece al mismo usuario de la imagen anchor.
- Imagen Negativa (N): Imagen que pertenece a distinto usuario de las imágenes anchor y positiva.

(a) Imagen anchor (b) Imagen positiva (c) Imagen negativa

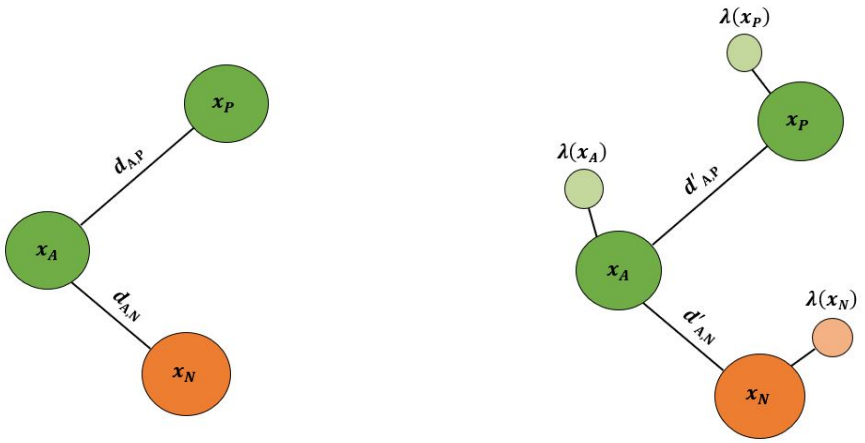
Figura 3.2: Ejemplo de imágenes que pueden formar un conjunto de Triplet Loss. [20]

Los triplets no serán aleatorios, sino que serán aquellos que cumplan la condición 3.1. El índice i indexa el triplet y $\|\cdot\|$ representa la distancia euclídea. El valor α es un número real que funcionará como umbral. X_A, X_P, X_N son los vectores de características resultado de pasar cada una de las imágenes anchor, positiva y negativa, por una arquitectura pre-entrenada para reconocer caras.

$$\left\|X_A^i - X_P^i\right\|^2 - \left\|X_A^i - X_N^i\right\|^2 > \alpha \tag{3.1}$$

El umbral α hará que los triplets se compliquen, ya que existirá una mayor distancia entre las imágenes anchor y positiva que entre las imágenes anchor y negativa. En otras palabras, esto significa que existirá mayor diferencia entre las imágenes de una misma persona que entre las imágenes de dos personas diferentes. Por lo tanto, el rendimiento será optimizado mediante la modificación de dichas distancias para conseguir el efecto contrario. En la Figura 3.3(a) se representa la estructura comentada. Se observa la estructura de los triplets sin tener en cuenta la información sensible de cada una de las clases. Con la finalidad de eliminar el sesgo, se añade a esta representación la información que se desea reducir. El resultado final se representa en la Figura 3.3(b).

Por todo ello se busca la generalización de la ecuación 3.1 para conseguir que $d_{A,N} > d_{A,P}$, mejorando así el rendimiento del algoritmo, al mismo tiempo que se reduce lo máximo posible $\lambda(X)$, función que representa el nivel de información sensible en la representación generada. Con este objetivo se propone la generación de una función de proyección del vector de características, dando lugar a una representación denominada agnóstica. De forma matemática la función de proyección se presenta mediante $\varphi(X)$. De esta forma se llega a una nueva función de perdidas donde se tiene en cuenta este nuevo planteamiento. Dicha función se indica en la ecuación 3.2.



(a) Estructura original de un triplet. [1]

(b) Estructura original de un triplet junto con la información sensible. [1]

Figura 3.3: La Figura muestra la estructura un triplet donde se puede ver que la distancia entre la imagen anchor y positiva es mayor que la que existe entre imagen anchor y negativa. En a) se ilustra la estructura original mientras que en b) se muestra la estructura original junto con la información sensible $\lambda(X)$. [1]

$$loss_1 = \sum_{i \in T} [\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2 + \Lambda^i] \tag{3.2}$$

Donde $[\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2$ representa la diferencia entre la distancia entre los vectores de características de las imágenes anchor y positiva y la distancia entre los vectores de las imágenes anchor y negativa. De esta forma se busca maximizar la distancia entre la imagen anchor y negativa y minimizar la distancia entre la imagen anchor y positiva. Asimismo, Λ^i representa la información sensible y se calcula siguiendo la ecuación 3.3. Por tanto, la ecuación puede dividirse en dos partes principales donde la diferencia entre las distancias busca la optimización de reconocimiento facial y la segunda parte, Λ^i, busca la reducción de la información sensible.

$$\Lambda^i = \lambda(\varphi(X_A^i)) + \lambda(\varphi(X_P^i)) + \lambda(\varphi(X_N^i)) + \alpha \tag{3.3}$$

Donde:

$$\lambda(\varphi(X)) = \log(1 + \sum_{s=1}^{C} \left| \frac{1}{C} - P_s(\varphi(X)) \right|) \tag{3.4}$$

Se propone eliminar la información sensible haciendo que la asignación de un usuario a cada una de las clases sea aleatoria. Por ello, C representa el número de clases y $\frac{1}{C}$ indica cual es la probabilidad

considerada aleatoria de pertenecer a una clase. Para género la probabilidad buscada será $\frac{1}{2} = 0{,}5$ y para etnia será $\frac{1}{3} \approx 0{,}333$. Por lo tanto S indica las clases existentes. $P_s(\varphi(X))$ representa la salida del modelo de clasificación de información sensible. Esta salida se representa matemáticamente como la probabilidad P$(S|\varphi(X))$. Por lo tanto, lo que se busca con esta ecuación es aproximar lo máximo posible la probabilidad de pertenecer a una clase a la probabilidad aleatoria. En otras palabras, la perdida será nula cuando el detector de información sensible tenga como salida $\frac{1}{C}$. El esquema final del modelo propuesto se representa la Figura 3.4. Para obtener la función de proyección $\varphi(X)$ se va a seguir el esquema de la Figura 3.5 el cual se explica por pasos a continuación.

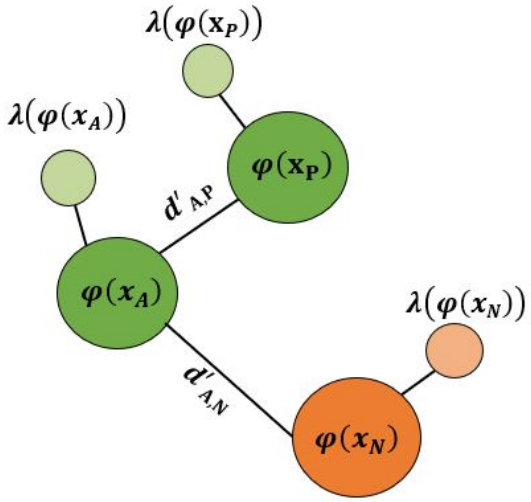


Figura 3.4: Representación agnóstica. Tras el entrenamiento del modelo propuesto, $d'_{A,P}$ se hace menor que $d'_{A,N}$ mientras que la información sensible también queda reducida $\lambda(\varphi(X)) < \lambda(X)$. [1]

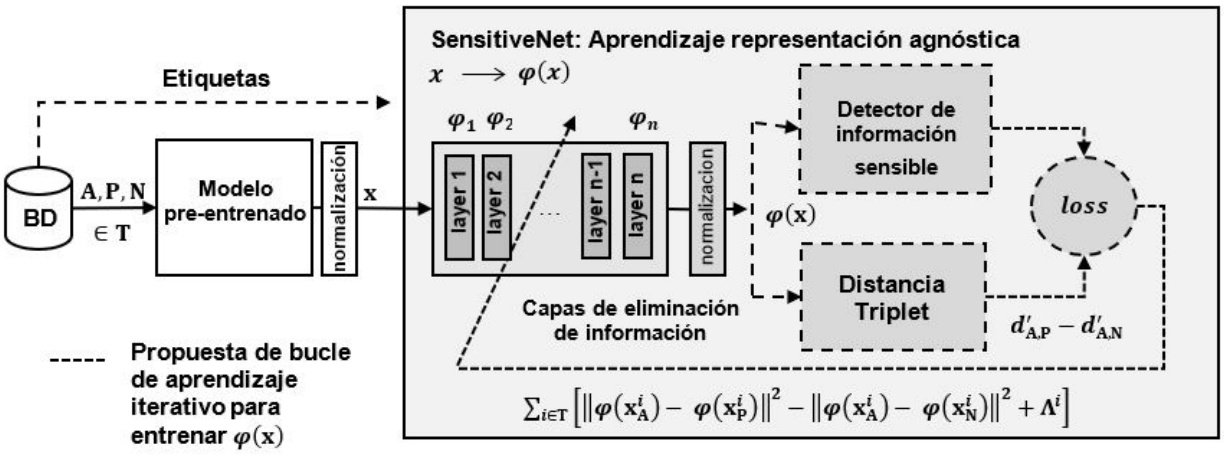


Figura 3.5: Proceso de eliminación de la información sensible de la representación del vector de características X. [1]

Paso 1: En primer lugar se debe obtener un vector de características. Este es generado por una

red pre-entrenada a partir de los triplets obtenidos. El vector X se obtiene tras una normalización de norma $L2$. Este paso inicial corresponde al bloque que se puede ver en la Figura 3.6.

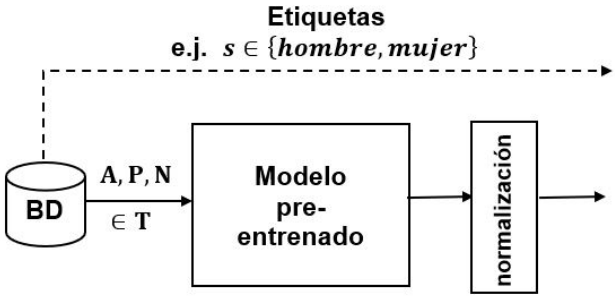


Figura 3.6: Proceso de obtención el vector de características inicial X a partir de una arquitectura pre-entrenada. [1]

Paso 2: A continuación, el detector $P(X)$ es entrenado a partir de X y la información sensible que queremos eliminar (etnia o género). Este modelo contiene una capa de clasificación densa con un número de unidades igual al número de clases C de la característica a eliminar. La función de activación es softmax. Por consiguiente, se genera, a partir del vector $X \in R^d$, una salida $P(X) \in R^c$. En otras palabras, el detector representará la probabilidad de que el rostro analizado pertenezca a una clase.

Paso 3: A continuación, se añade una capa densa conectada con l unidades y una función de activación lineal. De esta forma se llega a la obtención de $\varphi_1(X)$. Es decir, el vector X obtenido previamente se transforma en la proyección $\varphi(X)$.

Una vez obtenido $\varphi_1(X)$, el detector $P(X)$ es re-entrenado utilizando $\varphi_1(X)$ en vez de X. En la Figura 3.7 se representa el bloque comentado.

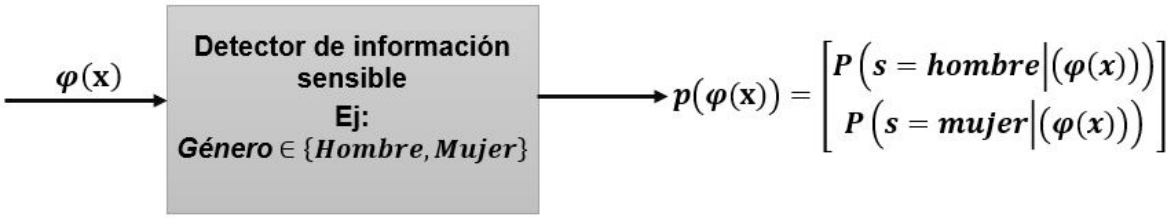


Figura 3.7: Representación del modulo de detección sensible. [1]

Paso 4: De forma paralela al clasificador sensible se encuentra un bloque que tiene como finalidad mejorar el rendimiento de reconocimiento. Es decir, se trata de un módulo basado en el algoritmo de triplet loss explicado previamente. Por lo tanto, tiene como objetivo reducir la distancia $d'_{A,P}$ mientras aumenta $d'_{A,N}$. La Figura 3.8 muestra este módulo.

Figura 3.8: Representación del módulo de optimización de rendimiento de verificación. [1]

A partir de ambos entrenamientos se llega a la ecuación de perdidas común representada en la ecuación 3.2. Por consiguiente, a partir de los pasos anteriores, $\varphi_1(X)$ es entrenada para reducir dicha función de perdidas a partir de los triplets de la lista T.

Los pasos 2-4 se repiten de forma iterativa n veces. Cada vez que se añade una capa se genera $\varphi_n(\varphi_{n-1})$, donde n señala la iteración de ese instante. En cada iteración, el detector $P(X)$ y el módulo de triplet loss son entrenados con la nueva proyección del vector de características. De esta forma, en cada repetición del bucle se minimiza la ecuación 3.2. Por consiguiente, el resultado final será la generación de $\varphi(X)$, entrenado para eliminar la información sensible mientras se mantiene un rendimiento competitivo de reconocimiento.

3.1. DiveFace: Base de Datos para Entrenamiento en la Diversidad y Evaluación de Algoritmos de Reconocimiento Facial

En esta sección se describe la base de datos utilizada para el método explicado anteriormente. Su creación ha resultado de la colaboración entre este proyecto y el realizado por la alumna Berta Fernández de la Morena.

Para la generación de este conjunto de datos, después de estudiar la posibilidad de uso de conjuntos como CelebFaces Attributes Dataset (CelebA) [21], VGGface2 [22] y Labeled Faces in the Wild (LFW) [23], se optó por la utilización de MegaFace MF2 [24]. Como su propio nombre indica, MF2 es parte del conocido conjunto MegaFace. Es una de las bases de datos públicas de reconocimiento facial con mayor numero de identidades. Esta consta de 672k identidades distintas con distintas imágenes por cada una de ellas, resultando en una suma de 4.7 millones de imágenes faciales. Todas ellas incluyen sus *"Bounding boxes"* y han sido obtenidas a partir de Flickr Yahoo's Dataset [24].

DiveFace ha sido utilizada en busca de los objetivos comentados en este proyecto. Por esta razón se ha buscado la creación de una base datos balanceada, evitando los posibles sesgos que introducen aquellas no balanceadas. Por ende, para la generación de esta, se han utilizado procesos automáticos

para realizar una distribución equitativa de datos entre las distintas clases. A pesar del uso de dichos procesos, fue necesaria la supervisión y corrección manual de los errores producidos. Por ello, se puede decir que los procesos para su generación han sido semi-automáticos en vez de automáticos. Así, contiene información distribuida entre 2 clases para género y 3 clases para etnia, resultando en un total de 6 clases. De forma más concreta cada clase se ha separado de la siguiente forma.

Etnia: Según el origen ancestral, en el mundo existen numerosos grupos étnicos, más de 5K según diversos estudios [1]. Sin embargo, para simplificar la clasificación de este proyecto, dicha división se ha reducido a 3 grupos. Esto se debe a que las diferencias entre ciertos grupos son mínimas, lo que aumenta considerablemente la dificultad en su clasificación. Con el objetivo de facilitar la agrupación, se han elegido las siguientes clases, las cuales tienen claras diferencias entre ellas.

- Origen en Japón, China, Corea y otros países asiáticos.
- Origen en África Subsahariana, India, Bangladés, Bután, entre otros.
- Origen en Europa, Norte América y Latinoamérica .

Género: Se divide en hombres y mujeres.

Para cada clase de las anteriores existen 4K identidades distintas. Cada una de ellas tiene una media de 5.5 imágenes, con un mínimo de 3 por persona. Por lo tanto, DiveFace contiene un total de 24K identidades y más de 120K imágenes faciales [1]. Las imágenes contienen las variaciones comentadas en este trabajo previamente. Entre ellas se encuentran variaciones de pose, calidad, iluminación, expresiones,etc.

Como se ha mencionado, DiveFace ha sido utilizada para el entrenamiento de los métodos propuestos en la sección anterior. No obstante, se han utilizado otras bases de datos para verificar que los resultados obtenidos con DiveFace se pueden aplicar en otros conjuntos, es decir, que los métodos son capaces de generalizar los resultados. Las dos bases de datos que se han utilizado en busca de este objetivo son LFW y CelebA.

LFW: Es una base de datos generada con el objetivo de optimizar el estudio del reconocimiento facial. Esta consta de más de 13K imágenes obtenidas en internet. Cada una de ellas está etiquetada con el nombre de la persona que aparece, es decir, proporciona el etiquetado de identidades. Contiene 1680 identidades distintas con más de 2 fotos por cada una. Las caras fueron detectadas mediante el detector Viola-Jones [23].

CelebA: Está formada por más de 10K identidades y más de 200K imágenes de personas famosas en el mundo. Esta proporciona información no solo sobre la identidad sino que también de distintos atributos. Más concretamente ofrece información de 40 atributos por imagen [21]. Entre ellos, a diferencia de la etnia, se encuentra el género. Por esta razón, la etnia tuvo que ser etiquetada mediante el uso del detector COTS. Tanto LFW como celebA, contienen diversas variaciones.

4

Desarrollo

Como se ha comentado en el apartado de diseño, los pasos a seguir para la finalidad del proyecto se representan en la Figura 3.5. De la misma forma se parte del método propuesto en [1]. Para llevarlo a cabo se utiliza la base de datos DiveFace. Por un lado, la lista de triplets T es generada a partir de 21K usuarios diferentes con 3 imágenes por cada uno de ellos. Como ya se ha comentado en el desarrollo de la base de datos, los usuarios están balanceados tanto en género como en etnia. Al umbral α, de la ecuación 3.1, se le asigna un valor de 0.2. Por otro lado, el número de las unidades de la capa densa que se añade en cada iteración será $l = 1024$.

En primer lugar, los vectores de características utilizados en este proyecto se obtienen de las arquitecturas pre-entrenadas VGGFace y ResNet-50, consideradas parte del estado del arte actual. ResNet-50 se trata de una red neuronal convolucional cuya característica principal es el uso de conexiones residuales que permiten saltos entre capas, a diferencia de otras arquitecturas convencionales. Esta está formada por 50 capas y 41 millones de parámetros [1]. De forma contraria VGGFace es una red neuronal convolucional tradicional que carece de la posibilidad de saltos entre capas. El método propuesto para eliminar la información de los vectores obtenidos puede dividirse en dos tareas principales, a) Mantener el rendimiento de verificación y b) Eliminar información sensible. A continuación se presenta como se ha llevado a acabo el desarrollo de ambas.

a) Mantener el rendimiento de verificación: Para esta tarea se van a utilizar dos bases de datos. Por un lado, la base de datos DiveFace va a ser utilizada en el entrenamiento. Por otro lado, se comprueba el rendimiento en la base de datos LFW, por ser un referente en la literatura y para evaluar la capacidad de generalización del modelo entrenado. Este proceso se realiza antes y después de eliminar la información sensible, lo cual permite observar cual es la perdida de rendimiento tras el proceso. En otras palabras, el análisis del rendimiento de verificación se realiza sobre los vectores X y $\varphi(X)$.

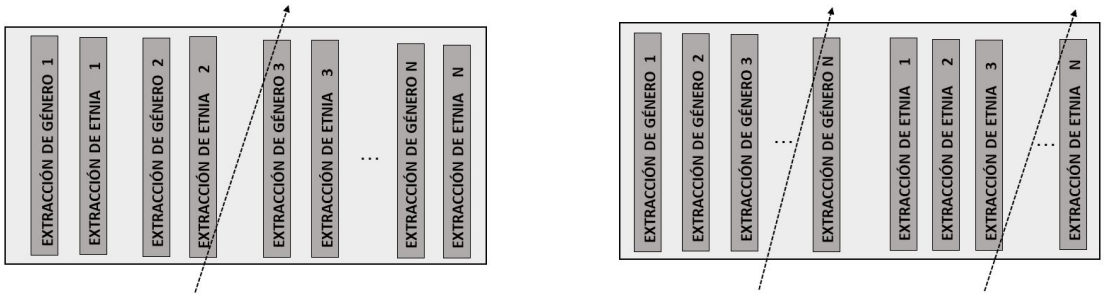
b) Eliminación de información sensible: Siguiendo el esquema 3.5 se entrena un modelo de clasificación de género o etnia en cada iteración del método. Por lo tanto, este es entrenado para el vector X y para cada $\varphi_n(X)$. De esta forma este proceso permitirá saber cual es la cantidad de información eliminada en cada iteración. Los modelos de clasificación de etnia y género son entrenados con la base de datos DiveFace. Para ello, esta se subdivide en imágenes diferentes para entrenamiento

y para test. La razón para usar únicamente DiveFace es que las bases de datos públicas actualmente tienen desbalanceo entre clases, lo que dificultaría notablemente el objetivo buscado. Por último, se evaluará el funcionamiento de la nueva representación en otra base de datos. En este caso se ha optado por analizar la generalización del algoritmo sobre CelebA.

Este proyecto propone el estudio de la extracción de la información en varios experimentos diferentes. En primer lugar, se propone la extracción de género y etnia de forma independiente. Una vez hecho esto, se propone el estudio final de la extracción tanto de etnia como de género al mismo tiempo. Debido a la necesidad de eliminar dos tipos de información sensible, el proceso explicado se puede realizar de dos formas diferentes, las cuales se explican a continuación.

Extracción intercalada: Este primer planteamiento propone la eliminación de género y etnia de forma alterna. Es decir, se propone un paralelizado de los entrenamientos de extracción de ambas informaciones en cada iteración del modelo. El proceso de evaluación del rendimiento de verificación se realiza tras eliminar ambos sesgos. La figura 4.1(a) ilustra la idea propuesta. En ella se puede observar como se van alternando cada una de las extracciones.

Extracción en serie: Este segundo planteamiento propone una mayor diferenciación entre tareas. Se plantea la extracción de un único tipo de información en cada iteración del método. De forma más concreta, este proyecto propone primero la eliminación toda la información de género y posteriormente la extracción de toda la información de etnia. La figura 4.1(b) ilustra esta propuesta. En ella se ilustra la serialización de la eliminación de género y etnia.



(a) Extracción intercalada

(b) Extracción en serie

Figura 4.1: La Figura muestra las técnicas de secuenciación de la extracción de información sensible propuestas. En a) se muestra la técnica de extracción intercalada. En b) se ilustra la técnica de extracción en serie.

Una vez definidas las diferentes formas de desarrollar el modelo planteado en [1], ha sido necesario realizar diversos experimentos cambiando diferentes parámetros del algoritmo para observar su efecto en los resultados. Para empezar, se ha modificado el valor de α, usado en la ecuación 3.3. El rango de variación que se ha planteado ha sido entre 0.1, 0.2 y 0.3. En cuanto a la ecuación de pérdidas propuesta en [1], se han planteado dos modificaciones principales, las cuales se indican en las

ecuaciones 4.1 Y 4.2.

$$loss_2 = \sum_{i \in T} [\log(1 + exp(\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2)) + \log(1 + exp(\Lambda^i)] \tag{4.1}$$

$$loss_3 = \sum_{i \in T} [\log(1 + exp(\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2)) + \Lambda^i] \tag{4.2}$$

En las tres funciones de perdidas comentadas, 3.2, 4.1 y 4.2, ha sido necesario modificar los pesos de cada una de sus partes explicadas en la sección de diseño. Así, se ha optado por añadir un factor de ponderación en cada una de ellas con la finalidad de evitar que la red reduzca únicamente una de ellas. Por lo tanto las ecuaciones resultan como se indica en 4.3, 4.4 y 4.5.

$$loss_1 = \sum_{i \in T} [(\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2) * \beta_1 + \Lambda^i * \beta_2] \tag{4.3}$$

$$loss_2 = \sum_{i \in T} [(\log(1 + exp(\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2))) * \beta_1 + (\log(1 + exp(\Lambda^i))) * \beta_2] \tag{4.4}$$

$$loss_3 = \sum_{i \in T} [(\log(1 + exp(\left\|\varphi(X_A^i) - \varphi(X_P^i)\right\|^2 - \left\|\varphi(X_A^i) - \varphi(X_N^i)\right\|^2))) * \beta_1 + \Lambda^i * \beta_2] \tag{4.5}$$

Siendo β_n el factor de ponderación comentado.

En relación a la evaluación del funcionamiento del algoritmo mediante clasificadores de género y etnia, se ha variado el número de épocas de entrenamiento de estos. De esta forma se puede observar si a pesar de aumentar este número, los clasificadores siguen siendo incapaces de realizar correctamente la clasificación.

Por último, se ha experimentado también con el número de capas a añadir, con la finalidad de observar la influencia de estas en el correcto funcionamiento del método. Asimismo, se ha modificado también el tipo de activación de la capa añadida en cada iteración. Más concretamente se han realizado experimentos con activación relu y activación lineal.

5 Integración, pruebas y resultados

5.1. Experimentación con ResNet-50

Este apartado muestra los resultados de los experimentos realizados con ResNet-50. La variable que indica la reducción de rendimiento se calcula como $Reducción = \frac{Antes - Después}{Antes - Probabilidad\ Aleatoria}$

Extracción de género

Tarea	Base de datos	Antes	Después	Reducción	Probabilidad Aleatoria
Identidad	LFW	98 %	95 %	6.25 %	50 %
Género	DiveFace	97.93 %	82.97 %	31.21 %	50 %

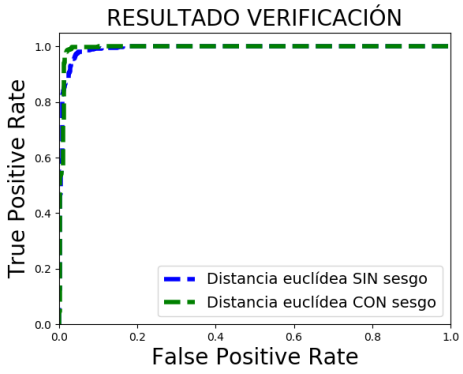
Tabla 5.1: Resultado del experimento de extracción de género. La tabla muestra los resultados del rendimiento en clasificación de identidad y género previa y posteriormente a la aplicación del modelo en extracción de género sobre la arquitectura de ResNet-50.

En la tabla 5.1 se pueden observar los resultados del algoritmo en la extracción de género. Estos muestran que tras aplicar el método se obtiene un rendimiento de clasificación de género del 82.97 %. En cuanto a la reducción del rendimiento de verificación se observa que la reducción es mínima, resultando este en un 95 %. Por lo tanto, el algoritmo es capaz de eliminar más de un 30 % de la información de género al mismo tiempo que se mantiene un alto acierto en verificación. En la Figura 5.1 se representan la curvas ROC de verificación y clasificación de género antes y después de la aplicación del algoritmo.

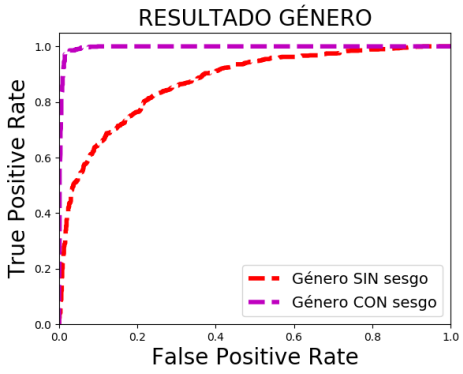
Extracción de etnia

Tarea	Base de datos	Antes	Después	Reducción	Probabilidad Aleatoria
Identidad	LFW	98.19 %	91.4 %	14.09 %	50 %
Etnia	DiveFace	97.54 %	53.34 %	68.83 %	33 %

Tabla 5.2: Resultado del experimento de extracción de etnia. La tabla muestra los resultados de rendimiento en clasificación de etnia e identidad antes y después de eliminar la información de etnia de los vectores obtenidos de la arquitectura de ResNet-50.



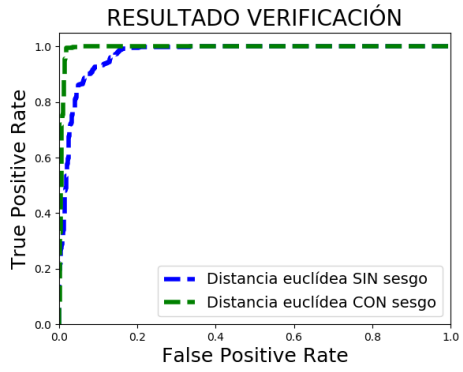
(a) Rendimiento de verificación antes y después del algoritmo.



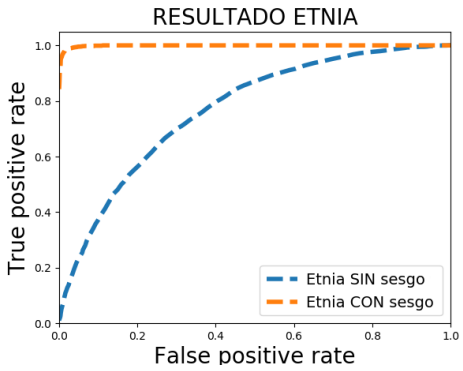
(b) Rendimiento de la clasificación de género antes y después del algoritmo.

Figura 5.1: Resultado del experimento de extracción de género. La Figura ilustra el rendimiento de a)verificación y b)género con los vectores obtenidos con ResNet-50, con sesgo y sin sesgo.

En la tabla 5.2 se pueden observar los resultados del algoritmo en la extracción de etnia. Estos muestran un buen funcionamiento del método en su eliminación. La reducción resulta en un 68.83 %, mientras que el rendimiento de reconocimiento no se ve afectado de forma notable, ya que este mantiene un 91.4 % de acierto. Las Figuras 5.2(a) y 5.2(b) muestran el rendimiento en verificación y en clasificación de etnia, respectivamente, antes y después del algoritmo.



(a) Rendimiento de clasificación de identidad.



(b) Rendimiento de clasificación de etnia

Figura 5.2: Resultado del experimento de extracción de etnia. La Figura ilustra las curvas ROC del rendimiento en clasificación de identidad y de etnia antes y después de la aplicación del algoritmo sobre los vectores de ResNet-50.

Extracción combinada

Tras la experimentación realizada en la investigación previa, se propone en análisis de la extracción tanto de género como de etnia de forma conjunta. Como se ha comentado en el apartado de diseño, se proponen dos formas diferenciadas de realizar dicha extracción. Por un lado se propone a) una extracción intercalada y por otro lado se propone b) una extracción en serie.

Extracción intercalada

Tarea	Base de datos	**Antes**	**Después**	**Reducción**	**Probabilidad Aleatoria**
Identidad	LFW	98 %	95.99 %	4.18 %	50 %
Etnia	DiveFace	97.93 %	35.04 %	97.35 %	33 %
Género	DiveFace	97.65 %	63.95 %	70.72 %	50 %

Tabla 5.3: Resultado del experimento de extracción intercalada. La tabla muestra el rendimiento de cada tarea de antes y después de la aplicación del algoritmo con extracción intercalada sobre los vectores de ResNet-50.

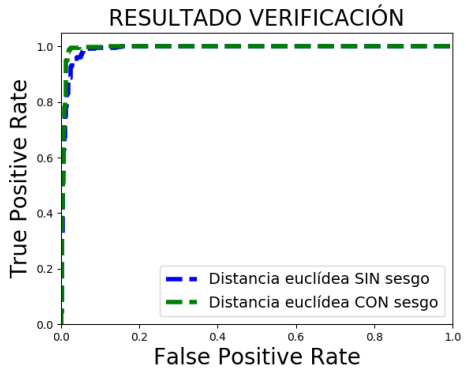
La tabla 5.3 muestra los resultados de la extracción intercalada. El rendimiento de las clasificaciones sensibles son ambos reducidos en más de un 70 %, mientras que el rendimiento de verificación se mantiene por encima de un 95 %. En comparación con los resultados de extracción de género, se observa una diferencia destacable. Este mostraba que el algoritmo solo era capaz de eliminar alrededor de un 30 % de información de género. No obstante, el experimento actual ha conseguido reducirlo en más de un 70 %. Por ello se puede deducir que la eliminación de información de etnia podría conllevar también la reducción de información de género. Estos resultados demuestran el éxito del modelo planteado en la eliminación de la información sensible (género y etnia en este caso) de los vectores de características. La Figuras 5.3(a), 5.3(c) y 5.3(e) muestran las curvas ROC del rendimiento de verificación, género y etnia, respectivamente, antes y después de la aplicación del algoritmo.

Extracción en serie

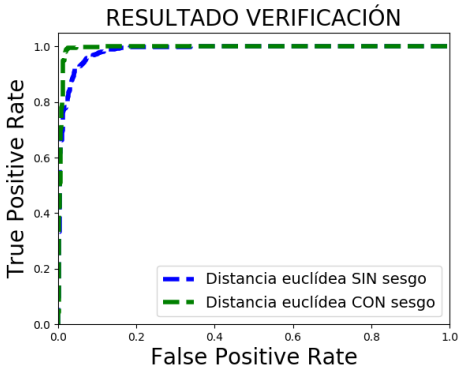
Tarea	**Base de datos**	**Antes**	**Después**	**Reducción**	**Probabilidad Aleatoria**
Identidad	LFW	98 %	93.6 %	9.16 %	50 %
Etnia	DiveFace	97.93 %	83.76 %	21.93 %	33 %
Género	DiveFace	97.65 %	77.23 %	42.85 %	50 %

Tabla 5.4: Resultado del experimento de extracción en serie. La tabla muestra el rendimiento de las tres tareas antes y después de la aplicación del algoritmo mediante la técnica de extracción en serie, sobre los vectores de ResNet-50.

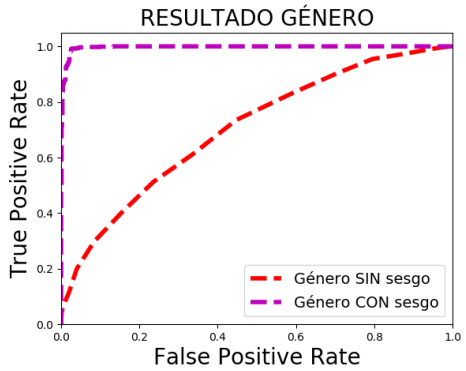
La tabla 5.4 muestra resultados menos óptimos que el caso anterior. Se observa que de esta forma el método no es capaz de eliminar más del 50 % de información de ninguna clase. La clasificación de etnia se reduce en un 21.93 % y la de género en un 42.85 %. Anteriormente se ha analizado la posibilidad de que la eliminación de información de etnia ayuda a la reducción de género, no obstante, este experimento propone un análisis más independiente entre clases. Esta contradicción podría ser la razón del empeoramiento de los resultados. Las Figuras 5.3(b), 5.3(d) y 5.3(f) muestran las curvas ROC del rendimiento de verificación, género y etnia, respectivamente, antes y después de la aplicación del algoritmo.



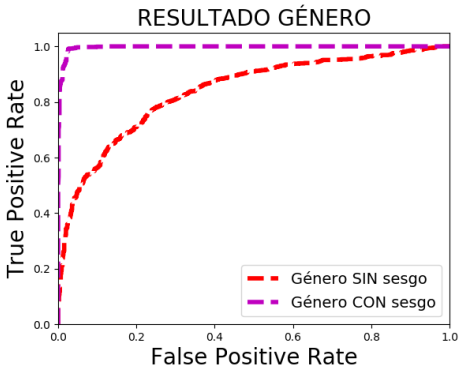
(a) Rendimiento de verificación de identidad antes y después del algoritmo de extracción intercalada



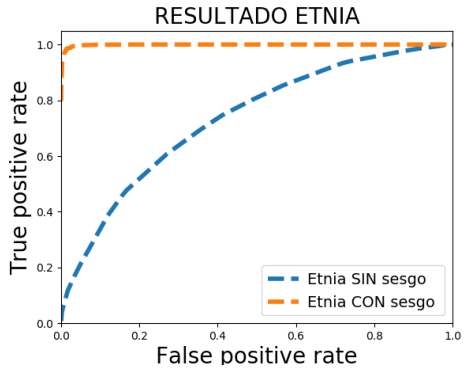
(b) Rendimiento de verificación de identidad antes y después del algoritmo de extracción en serie.



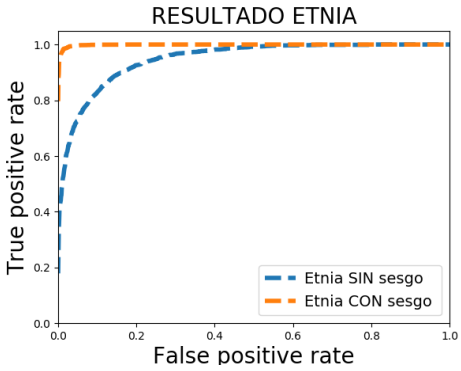
(c) Rendimiento de clasificación de género antes y después del algoritmo de extracción intercalada



(d) Rendimiento de clasificación de género antes y después del algoritmo de extracción en serie.



(e) Rendimiento de clasificación de etnia antes y después del algoritmo de extracción intercalada.



(f) Rendimiento de clasificación de etnia antes y después del algoritmo de extracción en serie.

Figura 5.3: Resultado del experimento de extracción intercalada y en serie. La Figura ilustra las curvas ROC de cada tarea antes y después de la aplicación del algoritmo con extracción intercalada y en serie sobre los vectores de características de ResNet-50.

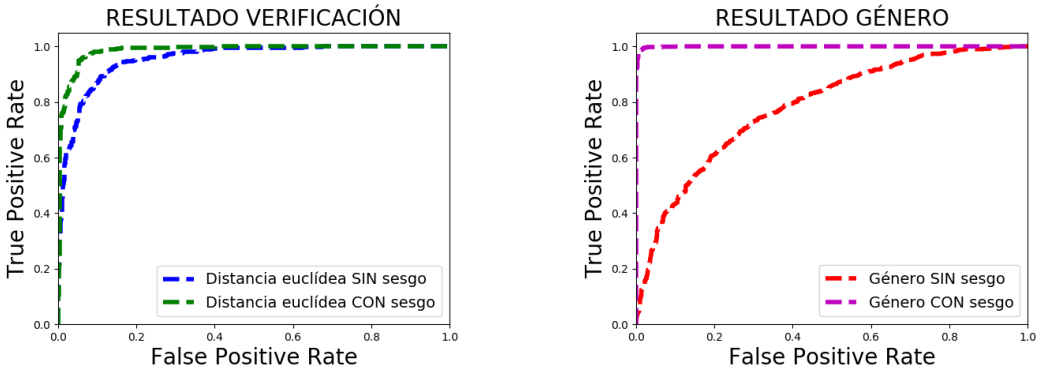
5.2. Experimentación con VGGFace

Este segundo apartado se enfoca en la eliminación del sesgo de los vectores de características obtenidos a partir de la arquitectura VGGFace. De forma más concreta, se experimenta en la extracción de información de género. El cálculo de la reducción se realiza de la misma forma que en la experimentación con ResNet-50.

Tarea	Antes	Después	Reducción	Probabilidad Aleatoria
Identidad	94.79 %	89.19 %	12.50 %	50 %
Género	98.99 %	71.03 %	57.07 %	50 %

Tabla 5.5: Resultados de la experimentación de extracción de género sobre los vectores de VGGFace. Los resultados muestran el rendimiento en las tareas de verificación y clasificación de género antes y después del algoritmo.

En la tabla 5.5 se pueden ver los resultados. En ella se puede observar que el algoritmo es capaz de eliminar bastante información de género, resultando en una reducción del 57.07 %. Este valor es mayor que el obtenido con ResNet-50, sin embargo, el rendimiento se ve más afectado, resultando su reducción en un 12.5 %. VGGFace basa su funcionamiento en una en una red lineal con mayor número de capas, mientras que ResNet-50, permite un funcionamiento recursivo. Este factor puede ser el motivo de las diferencias obtenidas entre ambas redes. En la Figura 5.4 se ilustran los resultados en la clasificación de identidad y género .



(a) Rendimiento de verificación antes y después del algoritmo.

(b) Rendimiento de clasificación de género antes y después del algoritmo.

Figura 5.4: Resultado del experimento de extracción de género. La Figura ilustra las curvas ROC de las tareas de verificación y clasificación de género antes y después de la aplicación del algoritmo en extracción de género sobre los vectores de características de VGGFace.

5.3. Evaluación del algoritmo

Este apartado experimenta la generalización del algoritmo en una nueva base de datos. En concreto, se experimenta la extracción de género en la base de datos pública CelebA. Para ello, se ha decidido evaluar el experimento con mejor resultado, el cual en este caso ha sido la extracción intercalada. El resultado obtenido ha sido un 49.1 % de rendimiento en clasificación de género, valor el cuál indica una reducción muy cercana al 100 %. La Figura 5.5 muestra el rendimiento en clasificación de género resultante.

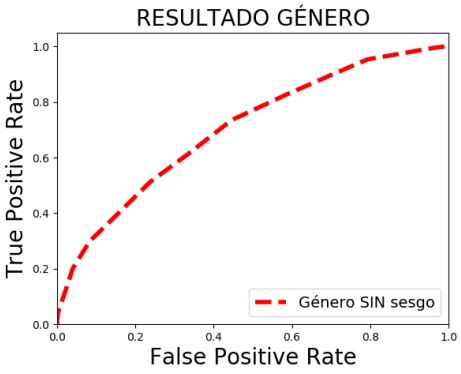


Figura 5.5: Resultado del rendimiento de clasificación de género sobre la base de datos celebA. La Figura muestra la curva ROC de la clasificación de género sobre CelebA tras la aplicación del algoritmo.

5.4. Parametrización

Los resultados mostrados en esta sección son fruto de numerosos experimentos que han permitido entender la influencia de cada parámetro en el proceso de aprendizaje. Para facilitar la lectura, no se han incluido todas las pruebas y se ha optado por seleccionar las más importantes. A continuación, la tabla 5.6 muestra la influencia que ha tenido cada parámetro a lo largo de toda la experimentación.

Parámetro	Descripción	Influencia
α	Parámetro de la ecuación 3.3	Media
Ecuación 3.2	Ecuación de perdidas usada por el algoritmo	Alta
β	Parámetro de ponderación de la ecuación de perdidas	Alta
Número de épocas	Épocas del entrenamiento de los clasificadores de género y etnia	Alta
Número de capas	Número de capas añadidas por el algoritmo	Media-baja
Activación	Tipo de activación de las capas añadidas por el algoritmo	Media-baja

Tabla 5.6: La tabla muestra el grado de influencia en los resultados de cada uno de los parámetros modificados en la fase de experimentación.

6

Conclusiones y trabajo futuro

6.1. Conclusiones

Actualmente las tecnologías basadas en biometría están cada vez más extendidas. Decisiones en diversos campos como el sector comercial, jurídico o médico, entre otros, cada vez confían más en los algoritmos de deep learning. Dichas decisiones pueden condicionar de forma notable la vida de una persona, por lo que hace plantearse a la sociedad la existencia de la discriminación algorítmica. Numerosos estudios han confirmado este problema, lo cual ha llevado a un aumento de la preocupación por el nivel de fiabilidad de las decisiones tomadas. Por todo ello, este proyecto ha planteado la generación de un método de reconocimiento facial entrenado para eliminar la información que puede causar este tipo de discriminación.

Este proyecto ha planteado un nuevo método entrenado para la eliminación de información sensible de la toma de decisiones de los algoritmos de aprendizaje profundo. Si bien este método puede ser aplicable en diversos campos de reconocimiento de patrones, este proyecto se ha centrado en el reconocimiento facial. Como ya se ha visto, la imagen de una cara puede dar una gran cantidad de información ampliamente relacionada con la discriminación en nuestra sociedad. Edad, género o etnia son algunas de la más controvertidas. Por este motivo, la investigación se ha enfocado en eliminar la información de género y etnia.

A lo largo de los últimos años se ha demostrado la gran influencia de las bases de datos en la aparición de la discriminación. Por esta razón, se ha optado, de forma adicional, por la creación de un nuevo conjunto de datos, DiveFace, balanceado en todas las clases. Con estas características, su creación ha facilitado notablemente el entrenamiento del modelo propuesto.

Para la generación del algoritmo, se propone un método iterativo centrado en una generalización de triplet loss. Se propone añadir la información sensible a la función de perdidas original, haciendo que esta no solo optimice el reconocimiento, sino que también introduzca la eliminación del sesgo. Los resultados han demostrado una gran capacidad del modelo para eliminar, al mismo tiempo, la información de género y etnia, mientras se ha mantenido un rendimiento competitivo. De forma más concreta, se han obtenido reducciones de más del 70 % de sesgo, manteniendo más de un 90 % de

acierto en verificación. Se concluye también que la eliminación de etnia podría contribuir notablemente a la eliminación de género. Por último, la comparación entre ResNet-50 y VGGFace ilustra una mayor capacidad que ResNet-50 de obtener información de género.

Actualmente la tecnología avanza notablemente obligando a la sociedad a avanzar de forma paralela a esta. Hace unos años era impensable la delegación de decisiones a los ordenadores. Actualmente, esta situación, que parecía imposible, es cada vez más normal. La tecnología tiene un objetivo claro y es mejorar la vida humana. No obstante, como se ha visto en este trabajo, una decisión sesgada puede dificultar las condiciones de vida de una persona por la única razón de pertenecer a un grupo vulnerable. Por ello, este trabajo busca la creación de un mundo algorítmico justo, donde el objetivo de la tecnología de ayudar a la sociedad se cumpla en cualquier situación y a favor de cualquier tipo de persona.

6.2. Trabajo futuro

Este proyecto se ha centrado en la eliminación de género y etnia. No obstante, existen diversas características que pueden dar lugar a la discriminación algorítmica. Por ello, como trabajo futuro, se plantea la eliminación de más características sensibles dentro del reconocimiento facial como, por ejemplo, la edad. Asimismo, resultaría interesante analizar el funcionamiento de los algoritmos propuestos sobre diversas bases de datos públicas y así poder observar cual de ellas dificulta más la extracción de información sensible. Sería de gran interés también analizar el funcionamiento del método con distintos tipos de clasificadores de información sensible. De esta forma se podría averiguar si el algoritmo es capaz de cegar la información a cualquier tipo de clasificador. Por último, se considera de gran importancia la realización de una evaluación del modelo en diferentes ámbitos del reconocimiento de patrones para observar su funcionamiento fuera del reconocimiento facial.

Bibliografía

[1] A.Morales, J. Fierrez, and R. Vera-Rodriguez, "Sensitivenets: Learning agnostic representations with application to face recognition," 2019.

[2] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork., "Learning fair representations," *Proc. of the Int. Conf. on Machine Learning, Atlanta, USA*, pp. 325–333, 2013.

[3] J. Buolamwini and T. Gebru., "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proc. of the ACM Conf.on Fairness, Accountability, and Transparency, New York, USA*, vol. 81, pp. 1–15, 2018.

[4] R.Oriondo, "Limitations of deep learning in ai research," 2019. Recuperado de https://medium.com/towards-artificial-intelligence/limitations-of-deep-learning-in-ai-research-5eed166a4205.

[5] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, and J. Fierrez., "Easuring the gender and ethnicity bias in deep models for face recognition," *Proc. of IAPR Iberoamerican Congress on Pattern Recognition, Madrid, Spain*, 2018.

[6] k. Hao, "This is how ai bias really happens—and why it's so hard to fix," *MIT Technology Review*, 2019.

[7] B. Goodman and F. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, 2016.

[8] M. Alvi, A. Zisserman, and C. Nellaker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," *Proc. of European Conf. on Computer Vision, Munich*, 2018.

[9] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, p. 671, 2016.

[10] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[11] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. Chen, V. Patel, C. Castillo, and R.Chellappa, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Processing Magazine*, vol. 35, pp. 66–83, 2018.

[12] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," *in Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 3476—-3483, 2013.

[13] A. Kumar, R. Ranjan, V. Patel, and R. Chellappa, "ace alignment by local deep descriptor regression," *arXiv preprint arXiv:1601.07950*, 2016.

[14] E. Raff and J. Sylvester, "Gradient reversal against discrimination," *Proc. of Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden*, 2018.

[15] S. Jia, T. Lansdall-Welfare, and N. Cristianini, “Right for the right reason: Training agnostic networks,” *Proc. of Int. Symposium on Intelligent Data Analysis, Hertogenbosch, Netherlands*, pp. 164–174, 2018.

[16] S. Hajian, J. Domingo-Ferrer, and A. Martinez-Ballester, “Discrimination prevention in data mining for intrusion and crime detection,” *Proc. of IEEE Symposium on Computational Intelligence in Cyber Security, Paris, France*, 2011.

[17] T. Kehrenberg, Z. Chen, and N. Quadrianto, “Interpretable fairness via target labels in gaussian process models,” 2018.

[18] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, “Auditing algorithms: Research methods for detecting discrimination on internet platforms,” *Proc. of the Annual Meeting of the International Communication Association, Seattle, USA*, 2014.

[19] A. Das, A. Dantcheva, and F. Bremond, “Mitigating bias in gender, age, and ethnicity classification: a multi-task convolution neural network approach,” *Proc. of European Conf. on Computer Vision Workshops, Munich, Germany*, 2018.

[20] O. Moindrot, “Triplet loss and online triplet mining in tensorflow,” 2018. Recuperado de https://omoindrot.github.io/triplet-loss.

[21] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From facial parts responses to face detection: A deep learning approac,” *Proc. of IEEE Int. Conf. on Computer Vision, Santiago, Chile*, pp. 3676–3684, 2015.

[22] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “Vggface2: A dataset for recognising face across pose and age,” *Proc. Int. Conf. on Automatic Face and Gesture Recognition, Xian, China*, pp. 67–74, 2018.

[23] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, “Labeled faces in the wild: A survey,” *Advances in Face Detection and Facial Image Analysis, Michal Kawulok, M. Emre Celebi, and Bogdan Smolka eds., Springer*, pp. 189–248, 2016.

[24] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, “The megaface benchmark: 1 million faces for recognition at scale,” *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4873–4882, 2016.


UAM
UNIVERSIDAD AUTONOMA
DE MADRID